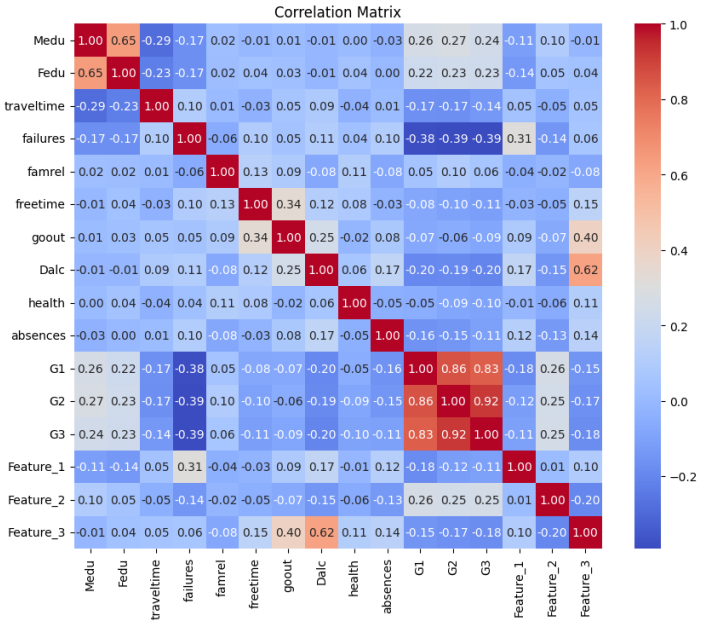
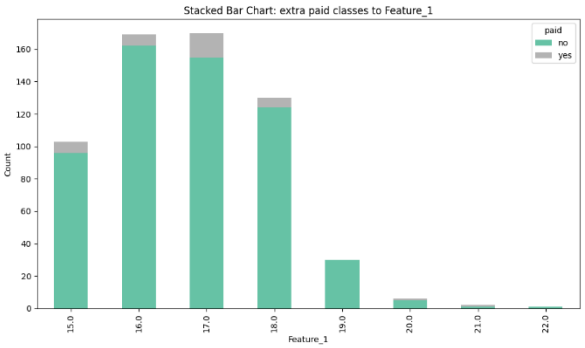
Task 1 : The CampusPulse Initiative

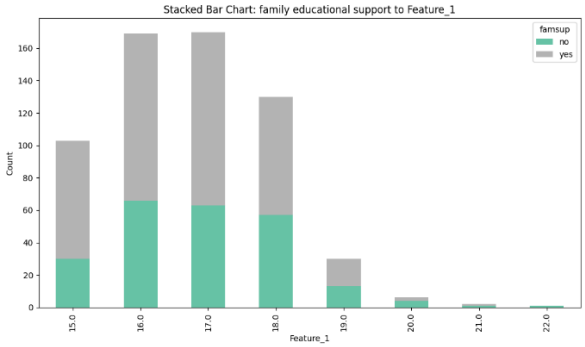
Level 1: Variable Identification Protocol

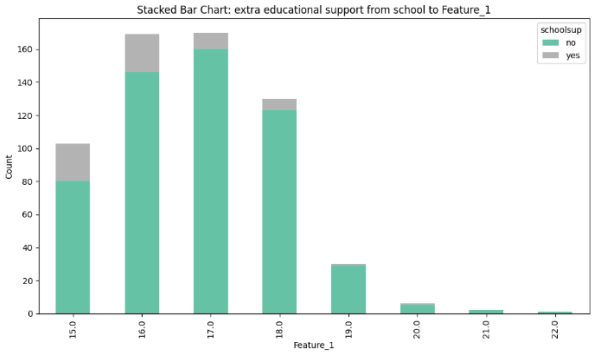


From the correlation matrix, we observe that Feature_2 has negative correlation with Feature_3. While Feature_1 has positive correlation with factors like number of past class failures, absences from school and also weekday alcohol consumption.

Feature_1:





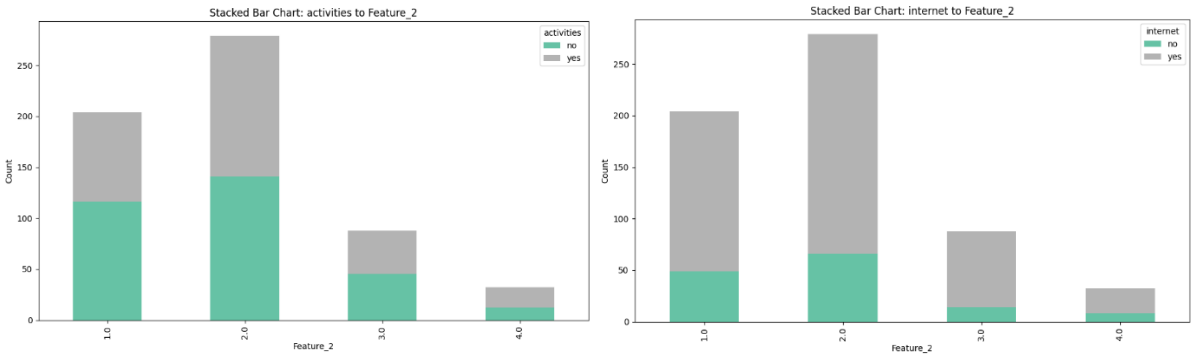


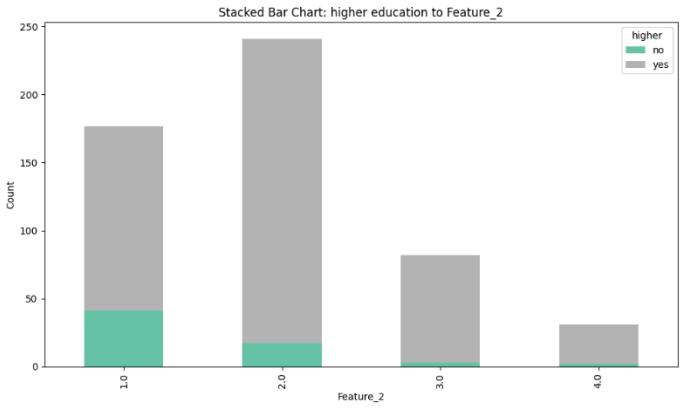
Feature_1	
15.0	103
16.0	169
17.0	170
18.0	130
19.0	30
20.0	6
21.0	2
22.0	1

- Plot 1 (Extra paid classes vs. Feature_1):
  Most students do not take extra paid classes, with only a small minority opting for them, and participation in extra classes drops as age increases.
- Plot 2 (Family educational support vs. Feature_1):
  Family educational support is most prevalent at smaller measures of Feature_1 (15–17) and declines as students get older.
- Plot 3 (School educational support vs. Feature_1):
  Extra educational support from school is uncommon, with the vast majority of students not receiving it, and support nearly vanishes for students having measure above 18.

Feature_1 values ranging from 15.0 to 22.0. These plots demonstrate that ***Feature_1 represents students' age*** by showing data concentrated in the age range typical for secondary education, and reveal that both family and institutional educational supports are more common at younger ages, tapering off as students grow older.

Feature_2:





Feature_2	
1.0	204
2.0	279
3.0	88
4.0	32

- In all three plots, Feature_2 is plotted on the x-axis with discrete, ordered values (1.0 to 4.0), and as Feature_2 increases, there is a clear trend toward more academically positive outcomes:
  - Higher Feature_2 values are associated with greater internet access and stronger aspirations for higher education.

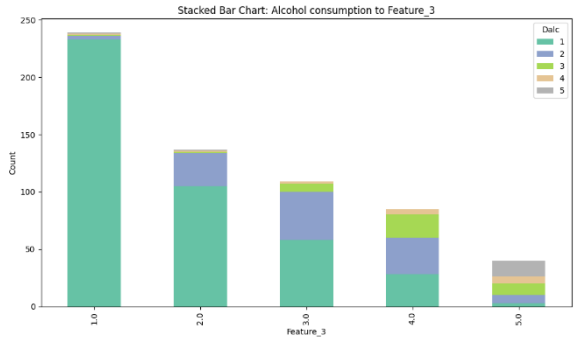
- The distribution of students shifts toward outcomes typically linked with higher academic achievement (e.g., more likely to want higher education, more likely to have internet access).

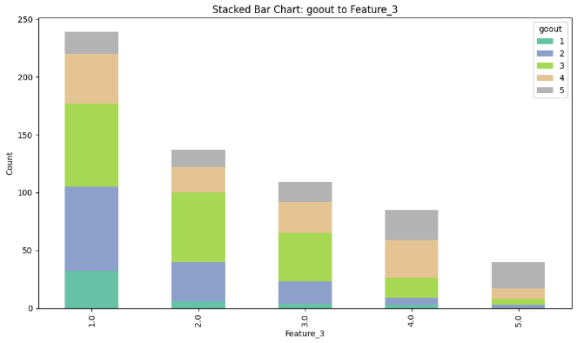
- The consistent association between higher Feature_2 and indicators of academic engagement and aspiration strongly suggests that ***Feature_2 is an academic metric***, likely representing academic performance or achievement level.

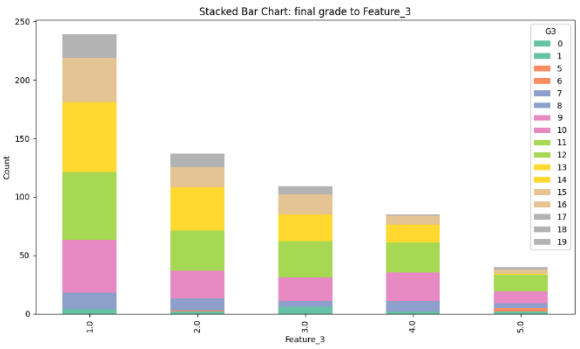
- The fact that higher Feature_2 values correspond to more students aspiring to higher education and having better access to academic resources is a classic pattern seen when plotting academic performance metrics

Thus, the structure and trends in these plots demonstrate that Feature_2 is closely tied to academic standing, with higher values reflecting greater academic achievement or engagement.

Feature_3:







Feature_3	
1.0	239
2.0	137
3.0	109
4.0	85
5.0	40

- Alcohol Consumption:
  The first plot shows a clear shift: at low Feature_3 values, nearly all students report minimal alcohol use, but as Feature_3 increases, higher levels of alcohol consumption become much more common. This demonstrates a strong positive correlation between Feature_3 and alcohol use.

- Social Activities:
  The second plot reveals that higher Feature_3 values correspond to more students reporting frequent social outings. The distribution shifts from low "goout" scores at

low Feature_3 to high "goout" scores at high Feature_3, confirming that Feature_3 is linked to social activity patterns.

- Academic Impact:
  The third plot shows that as Feature_3 increases, the distribution of final grades shifts downward, with fewer high-achieving students and more low-achieving students. This suggests that higher engagement in social activities and alcohol consumption (as captured by Feature_3) is associated with lower academic performance.

Thus**,** the stacked bar charts collectively demonstrate that ***Feature_3 is strongly correlated with both social activity levels and alcohol consumption patterns***, and that higher values of Feature_3 are linked to increased socialization, higher alcohol use, strong peer influence and lower academic achievement.

Level 2: Data Integrity Audit

No. of columns having empty spaces: 10

The columns having empty spaces are: 'famsize', 'Fedu', 'traveltime', 'higher', 'freetime', 'absences', 'G2', 'Feature_1', 'Feature_2', 'Feature_3'.

Feature-Specific Imputation Strategies:

For Categorical Variables:

- **Mode imputation**: Use most frequent value for categorical features with low missing percentages (<5%)

For Numerical Variables:

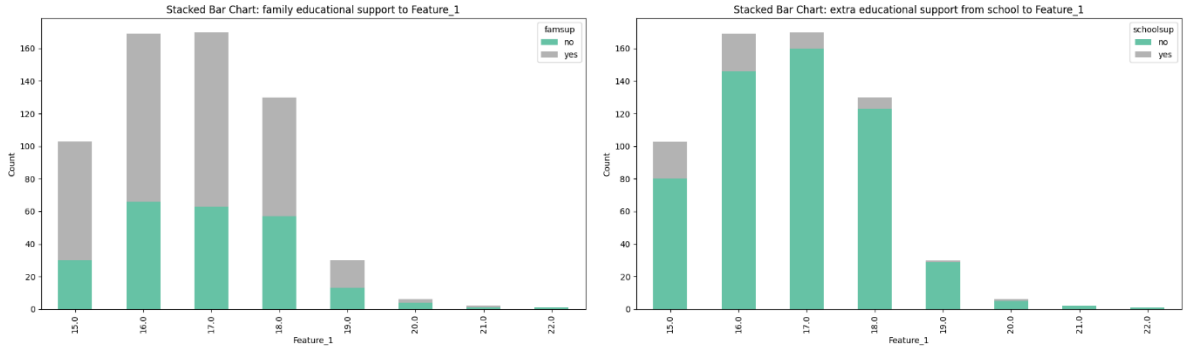
- **Mean imputation**: For normally distributed numerical features with minimal missing data
- **Median imputation**: For skewed distributions or presence of outliers

For Mixed Data Types:

- **Multiple Imputation by Chained Equations (MICE):** For complex datasets with multiple missing variables
- **KNN imputation**: Effective for both numerical and categorical mixed datasets

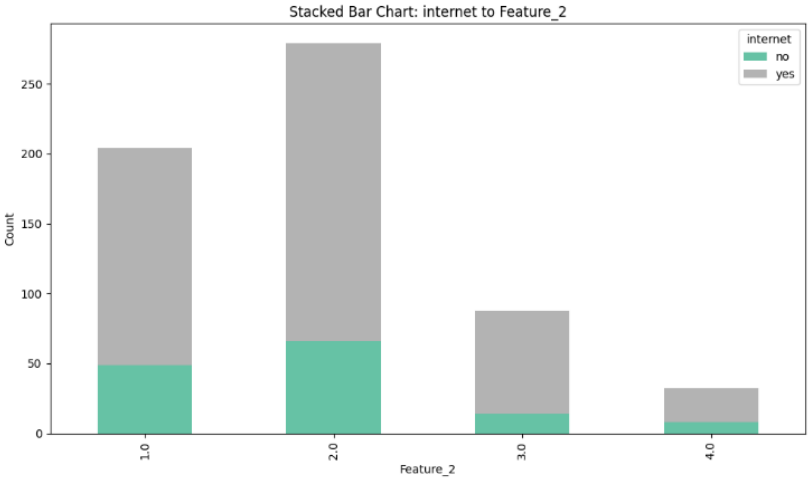
Level 3: Exploratory Insight Report

Question 1: How does age (Feature_1) distribution relate to educational support systems?



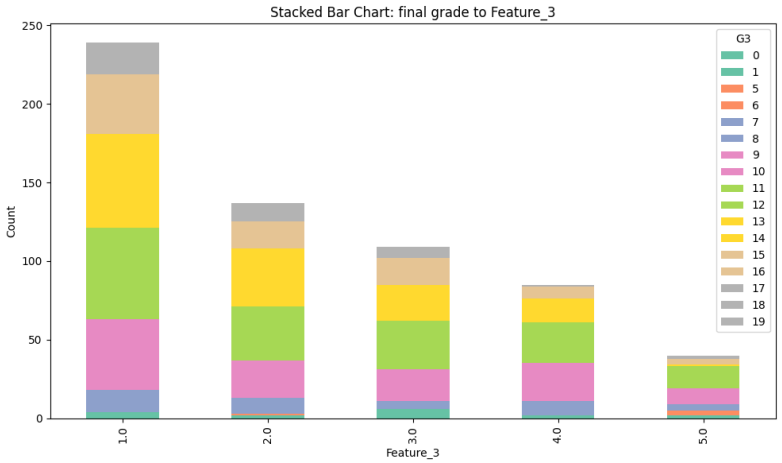
Insight: Students in the 16-17 age range appear to receive the highest levels of educational support across all categories. The visualization demonstrates that younger students (ages 15-16) show higher dependency on family educational support, while older students (ages 18+) exhibit decreased engagement with formal support systems. This pattern suggests that as students mature, they either become more independent in their learning or may require different types of academic intervention strategies

Question 2: What is the relationship between academic engagement (Feature_2) and technology access?



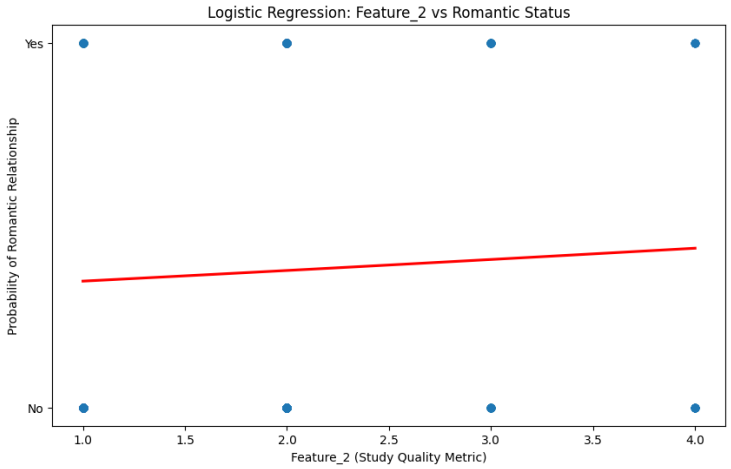
Insight: Students with Feature_2 values of 3-4 show significantly higher rates of internet access and extracurricular participation, suggesting that academic engagement extends beyond classroom performance to encompass broader educational experiences. This emphasizes the importance of digital literacy and extracurricular opportunities as complementary factors to traditional academic performance, suggesting that holistic educational approaches yield better outcomes than purely academic interventions.

Question 3: How do social behaviours and peer influence (Feature_3) impact academic outcomes?



Students with higher Feature_3 values demonstrate increased engagement in potentially risky social behaviours, including higher alcohol consumption and frequent social outings, which correlates with lower final grade distributions across the performance spectrum. Intensive peer influence often leads students toward social activities that compete with academic focus, creating a trade-off between social integration and academic achievement.

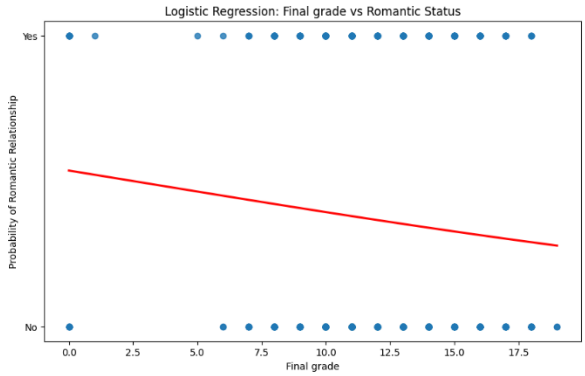
Question 4: How does academic engagement (Feature_2) influence romantic relationship likelihood?

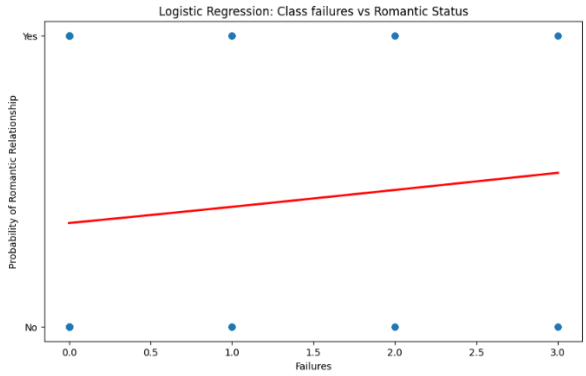


The logistic curve shows that higher academic engagement correlates with lower romantic relationship probability, it might suggest that academically focused students prioritize educational goals over social relationships. However, it is to be noted that as the regression coefficient, which corresponds to the slope of the regression is significantly small yet slightly positive, it could indicate that socially confident, engaged students successfully balance

multiple life domains and successfully manage their academics despite of being in a romantic relationship.

Question 5: How do academic failures predict romantic relationship engagement?



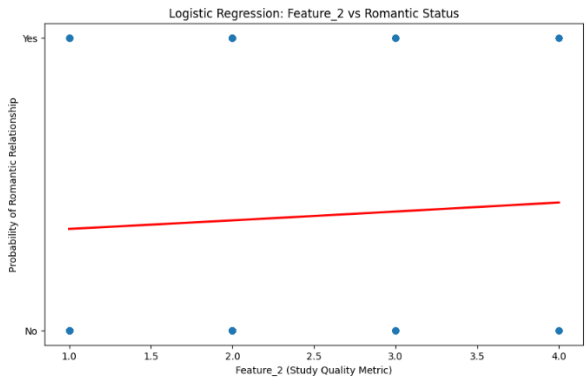


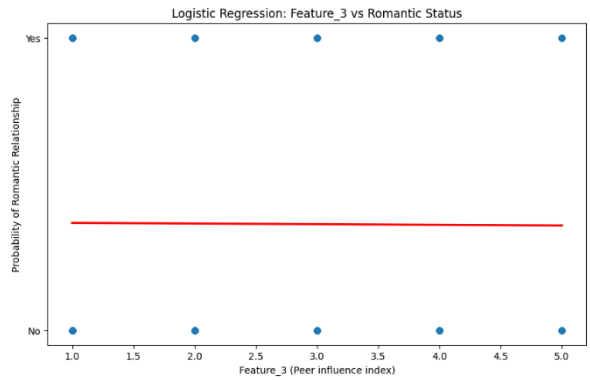
The visualization reveals whether students experiencing academic difficulties are more or less likely to engage in romantic relationships, potentially indicating whether relationships serve as emotional support during academic challenges or represent distractions from academic recovery efforts. Thus, academic failures correlate with increased romantic relationship likelihood, it might suggest that students seek emotional support and validation through personal relationships when facing academic challenges.

Level 4: Relationship Prediction Model

On using multiple models like **Logistic regression**, Random Forest classifier and lightBGM and comparing their accuracies it is observed that Logistic regression model is recommended as it has a greater accuracy of **68.21%** as compared to other models like Random Forest and LightBGM with 62.05% accuracy.

Plots obtained in Logistic regression:





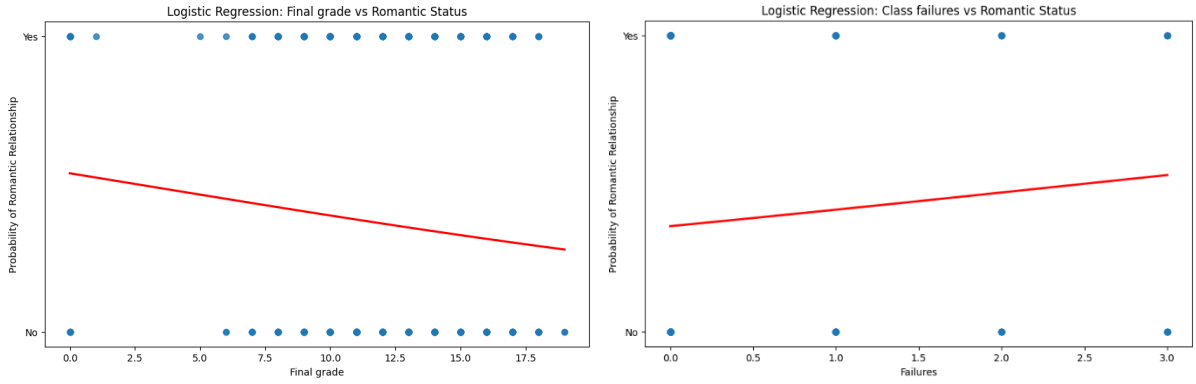
Plot of Feature_2 vs Romantic status:

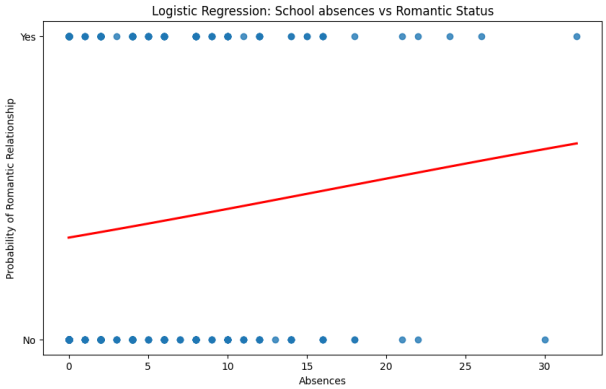
- The red logistic curve has a slight upward slope, suggesting a very weak positive relationship between academic performance and the probability of being in a romantic relationship.
- However, the trend is almost flat, indicating that Feature_2 (academics) does not strongly influence romantic status.
- At each academic level (1 through 4), students are both in and not in relationships.
- This implies that students across all academic performance levels show mixed romantic statuses.

Plot of Feature_3 vs Romantic status:

- The red line is perfectly flat, indicating no relationship between Feature_3 and the probability of being in a romantic relationship.
- At every value of Feature_3 (from 1 to 5), students are evenly split between being in and not in relationships.
- There is no visible trend or concentration of relationship status at any level of peer influence.

More plots about romantic status which uncovers what patterns in academic metrics data might signal a student's likelihood of being in a romantic relationship:

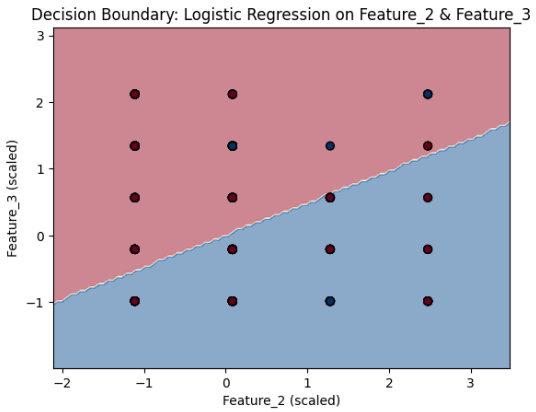




Level 5: Model Reasoning & Interpretation

Decision boundary plot

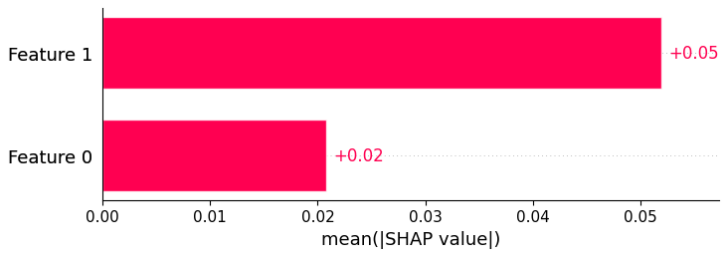
For plotting decision boundary, Logistic regression model is being used as it had greater accuracy of 68.21% as compared to other models like Random Forest and LightBGM with 62.05% accuracy.



- The diagonal boundary reveals that romantic relationship status depends on the combination of academic performance and social activity levels, rather than either factor independently.
- The boundary shows that both high-performing students with moderate social activities and moderately performing students with higher social engagement can be in relationships, challenging simplistic assumptions.
- Scattered data points across both regions emphasize that individual circumstances matter beyond these two metrics alone.

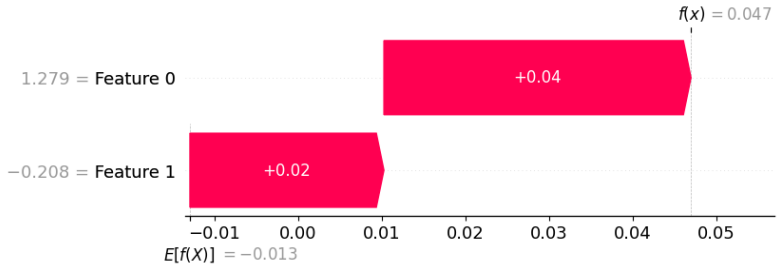
SHAP Analysis

(note: Feature 0 is equivalent to Feature_2; Feature 1 is equivalent to Feature_3)



Feature_3 (social activities/alcohol consumption) shows significantly higher overall importance (0.05) compared to Feature_2 (academic performance, 0.02), indicating that social behaviour patterns are 2.5 times more influential in predicting romantic relationship status than academic metrics.

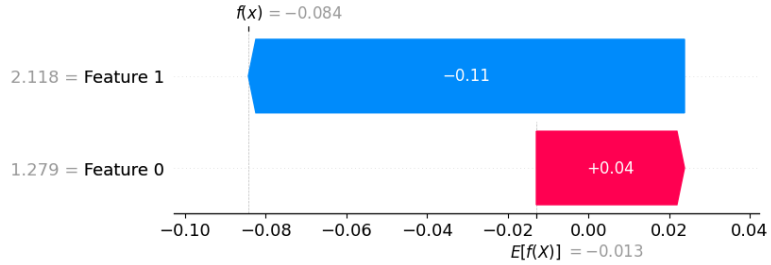
YES prediction



High academic performance (Feature_2 = 1.279) contributes +0.04 toward a "yes" prediction, supporting the idea that academically successful students can maintain relationships.

Lower social activity level (Feature_3 = -0.208) still contributes +0.02 positively, suggesting that moderate social engagement (rather than excessive partying/drinking) supports relationship formation.

NO prediction

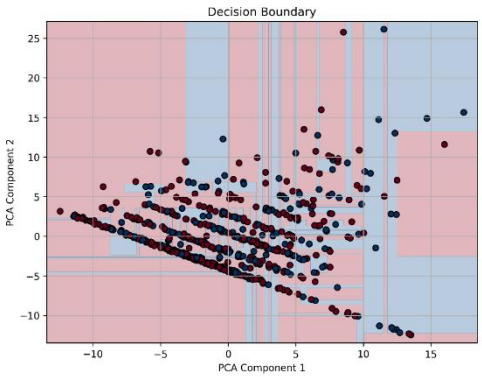


High Social Activity as Barrier: Very high social activity/alcohol consumption (Feature_3 = 2.118) creates a strong -0.11 negative influence, suggesting that excessive partying and drinking patterns may hinder relationship formation or stability.

Academic Performance Still Positive: Despite the "no" prediction, academic performance (Feature_2 = 1.279) still contributes +0.04 positively, indicating that academic success alone isn't sufficient to overcome very high social activity patterns.

Bonus Level: The Mystery Boundary Match

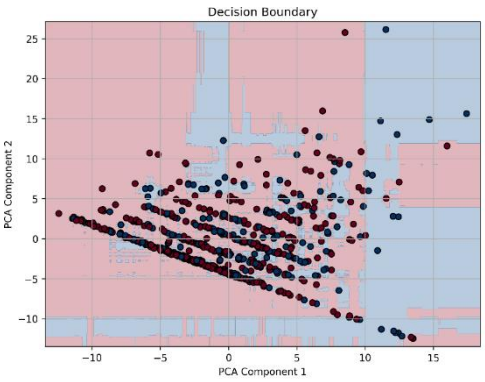
Plot-1:



Plot_1: Decision Tree Classifier

- Rectangular Partitioning
- Axis-Aligned Splits (orthogonal boundaries) decision boundaries aligned along axes
- Stepwise Complexity

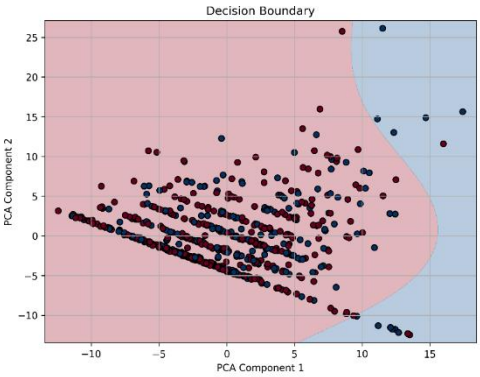
Plot -2:



Plot_2: K-Nearest Neighbors (KNN) Classifier

- tiny blue/red patches scattered like islands
- Smooth Curved Boundaries

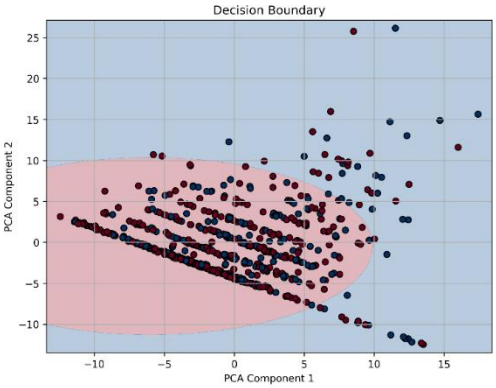
Plot -3:



Plot_3: Logistic Regression

- Smooth, linear-ish boundary (continuous, with a single smooth curve separating the two regions)
- No sharp partitioning
- Global linear decision function

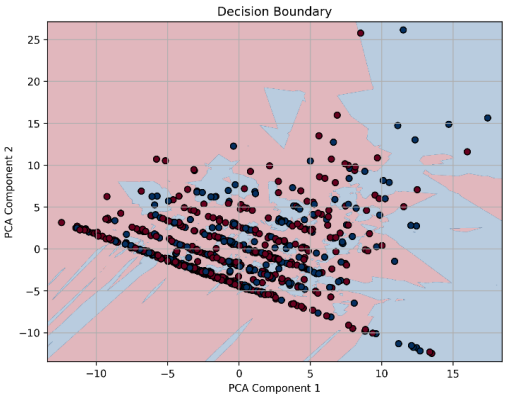
Plot -4:



Plot_4: SVM with RBF kernel

- Non-linear and smooth
- PCA components are linear transformations of the original features, but the non-linear boundary here suggests the original model operates non-linearly
- Linear models (e.g., logistic regression) would fail to capture the circular/curved separation visible in the plot.

Plot -5:



Plot_5: K-Nearest Neighbors (KNN) Classifier

- tiny blue/red patches scattered like islands
- Smooth Curved Boundaries & Non-Linear and Fragmented Regions
- Hard classification regions with abrupt changes rather than a soft gradient between classes